DOCUMENT RESUME

ED 191 917 TM 800 543

AUTHOR Cohen, Allan S.: Kappy, Kathleen A.
TITLE Achievement Level Testing Effects on Rasch Item Difficulty Estimates.
PUB DATE Apr 80
NOTE 15p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests: *Difficulty Level: Elementary Education: Item Analysis: Latent Trait Theory: Quantitative Tests: Reading Tests: Regression (Statistics): *Scores: *Test Items: *Test Norms
IDENTIFIERS *Out of Level Testing: *Rasch Model: Sequential Tests of Educational Progress

ABSTRACT

The ability of the Rasch model to provide item difficulties and achievement test scores which are invariant is studied. Data for the study were obtained from students in grades 3 through 7 who took the Sequential Tests of Educational Progress (STEP III) Reading and Mathematics Concepts tests during a spring norming study. Each test contained 50 items administered in a 40-minute period. These norming data were collected so as to provide both in-level and out-of-level norms, although only a portion of the full set of norming study data was used. All calibrations to the Rasch model were done using the BICAL program. Separate item difficulty and score scale estimates were obtained for each grade level administered a particular level of the tests. Results indicated that achievement scores and item difficulty parameters could be considered as essentially invariant. Ten of 14 out-of-level regressions of item difficulty estimates on in-level estimates were linear, of unit slope, and had zero intercepts. The remaining four regressions had zero intercepts and were linear but did not have slopes of 1.0. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED191917

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Achievement Level Testing Effects on Rasch Item Difficulty Estimates

Allan S. Cohen and Kathleen A. Kappy

University of Wisconsin-Madison Southern California Edison

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
A. Cohen
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM800543

Running Head: Achievement Level Effects

Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Rev. 3

Achievement Level Testing Effects on

Rasch Item Difficulty Estimates

Abstract

When students' achievement levels are different from those of their in-grade-level test, out-of-level testing is sometimes used to obtain more reliable estimates of their achievement. Providing empirical norms for out-of-level testing requires each level of the test be re-normed on appropriately representative norming samples. This additional norming is required since there is a dependency in classical test theory between the test score and the specific set of test items taken. A solution to the need for out-of-level norming may be to use the 1-parameter Rasch Model to provide item calibrations. Cost savings from this use of the Rasch Model would accrue from the presumed invariance of the item and score parameters. Use of the Rasch Model results in cost savings since the assumption of achievement score and item difficulty invariance with respect to grade level obviates the need for re-norming. This study attempted to determine whether the Rasch Model can be used to provide item difficulties and achievement test scores which are indeed invariant. The data used were obtained from data collected on in-level and out-of-level norming samples. Results indicated that achievement scores could be considered as essentially invariant. Item difficulty parameters also showed a high degree of invariance. Ten of 14 out-of-level regressions of item difficulty estimates on in-level estimates were clearly linear, of unit slope, and had zero intercepts. The remaining 4 regressions had zero intercepts and were linear but did not have slopes of 1.0. Misfitting items may have caused some of this apparent lack of invariance.

Achievement Level Testing Effects on Rasch Item Difficulty Estimates

Testing students' achievement at a grade level different from their own is referred to as "out-of-level" testing. The intention of out-of-level testing is to more closely match the difficulty of the test to the student's achievement level than would otherwise be the case using an in-level test form thereby increasing the reliability of the score estimate. Obtaining norms for such out-of-level testing using classical test theory, however, means that appropriately representative norming samples need to be administered the out-of-level form. While such additional norming does provide useful information otherwise unavailable using classical test theory methods, test norming costs rise markedly for each instance of an achievement test which is so re-normed.

The Rasch model (Rasch, 1960; Wright, 1977), however, may offer some important advantages over classical test theory procedures for obtaining out-of-level norming estimates. These advantages derive from one of the assumptions of the Rasch Model, that log item difficulties remain invariant over samples drawn from the same population. One objective of this study was to examine the validity of the invariance assumption for in-level and out-of-level estimates of Rasch item difficulties.

A second purpose of the study concerned achievement scores. According to classical test theory, the distribution of achievement scores is specific to (among other things) the population from which the norming sample was drawn as well as the set of test items administered to that sample. Such specificity with respect to items means that the norms for a test can be interpreted only in the context of the same items as administered to the

norming samples. There is, in other words, a dependency between the specific set of items on the test, the test score scale, and the distribution of test (or achievement) scores in the population. Therefore, if a test is to be used out-of-level, then norms for the test must also be obtained on an appropriate out-of-level sample.

This type of dependency is not present using the Rasch model. The same latent trait is measured on a Rasch scale whether one uses the full set or any subset of the set of items which fit the model for that scale (i.e., usually scale is taken to mean for that test). There is not, in other words, a dependency between a specific subset of items and the Rasch score scale. The only dependency in the norming context is between the score scale and the distribution of achievement scores in the population. The Rasch model further assumes that the log achievement score scale remains invariant over groups from the same population.

Using this reasoning, it should be possible to obtain an achievement level score that is the same as the in-grade score and which can be related to in-grade norms even though it may have been obtained from answers to out-of-level items. A second objective of this study, therefore, was to determine whether in-level and out-of-level Rasch score scales are indeed the same, i.e., invariant.

Methods

Data Source

Data for this study were obtained from examinees' responses collected for this spring norming study on the STEP III Reading and Mathematics Concepts tests (ETS, 1979). The STEP III Reading test measures vocabulary in the context of other material, literal comprehension, and interpretation of written material. The test items are written using academic and non-academic material including drama, poetry, narrative, newspaper articles and instructions.

The STEP III Mathematics Concepts test measures understanding of the decimal system and ability to recognize the patterns and properties of operations with whole numbers, fractions, and decimals, ability to recognize reasonable answers, and to solve problems or to recognize when there is insufficient information for a solution. At least thirty percent of the test involves problem solving; questions on measurement cover only the metric system; and geometry questions are included to some extent at each level. Each test contained 50 items administered in a 40-minute period.

These norming data were collected so as to provide both in-level and out-of-level norms according to classical test theory methods. That is, the samples of students were drawn to provide a nationally representative distribution of achievement levels for students at each of the grade levels covered by these tests. For purposes of this study, only a portion of the full set of norming study data was used (see Table 1).

- - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - -

Rasch Model Parameter Estimates

All calibrations to the Rasch model were done using the BICAL program (Wright & Mead, 1978). Separate item difficulty and score scale estimates were obtained for each grade level administered a particular level of the tests included in this study.

Results

Item Difficulties. To determine whether item difficulty estimates for in-level and out-of-level norming situations remained invariant, each set of Rasch item difficulty estimates was regressed onto the in-level estimates for that grade. Looking at Table 1 as an example, for Reading level G, grade 5 is the in-level grade. Estimates of the item difficulties obtained on a 4th grade sample (N = 909) for the 50 reading test items were regressed on those for the in-level fifth grade sample (N = 957). If the out-of-level estimates regress linearly with a slope of 1.0 and an intercept of zero onto the in-level estimates, then invariance is very likely present (Cook, Eigenor, Hutton, 1979; Hashway, 1979).

- - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - -

Summary data for these regression analyses (see Tables 2 a and b) indicate that the regressions were linear and had intercepts which were very likely zero.

For most of the regressions, the slopes of the regression lines were not different from 1.0. For the reading test, 6 of 7 slopes were 1.0; for the mathematics concepts tests, 4 of 7 slopes were 1.0. These four non-unit

slopes raised some concern primarily about fit of these items to the model. Therefore, further analysis was done to determine the fit of items to a Rasch scale. BICAL prints out a set of mean square fit statistics. One approach to take in determining fit is to calculate an upper value for these statistics beyond which items are summarily rejected. This upper value is computed as 1.0 + [3 x standard error of the mean square fit statistics distribution for that test].

Using this value to reject items from the regression analysis, however, did not result in any significant improvement in the four regression slopes. It may be that the BICAL fit estimate alone does not provide enough information to determine lack of fit to the model. A more thorough analysis of the items would seem to be a better procedure for determining fit. It would seem important, for instance, to look at (1) the content of each of the items, (2) the item difficulty values to determine the presence of outliers, and (3) to point biserial correlations.

Achievement Scores. With respect to achievement scores, it was expected that, since the item difficulty estimates were only moderately invariant (except as described above), the score scales would behave similarly.

- - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - -

The results in Table 3 indicate that the out-of-level regressions of achievement scores on in-level scores were all linear, had slopes not different from 1.0, with intercepts of zero. These results are consistent with previous results (Wright, 1968) and suggest that the achievement scores may be invariant whether computed on in-level or out-of-level samples.

Conclusions

The results of the regressions of out-of-level on in-level item difficulty estimates lend moderate support to the validity of the assumption of invariance. A further examination of the four non-unit regressions may provide better insight into the reasons for the lack of invariance. Based on the results of this study, it seems likely that out-of-level achievement score estimates will be similar to in-level estimates, at least for the types of tests and the grade levels measured in this study.

References

Cook, L. L., Eigenor, D. R., and Hutten, L. R. Considerations in the application of latent trait theory to objectives-based criterion-referenced test. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Educational Testing Service Sequential tests of educational progress, III. Palo Alto, California: Addison-Wesley, 1979.

Hashway, R. M. Individual and program evaluation using the Rasch model. New York: Praeger Publications, 1978.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Denmark Paedegogical Institute, 1960.

Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1968.

Wright, B. D. Misunderstanding the Rasch Model. Journal of Educational Measurement, 1977, 14(2), 219-225.

Wright, B. D. & Mead, R. J. BICAL: Calibrating rating scales with the Rasch model. Research memorandum 23A. Chicago, Illinois: University of Chicago, Department of Education Statistical Laboratory, 1976.

Table 1

STEP III Spring Norming Sample Sizes

Used for This Study

Grade Level	Reading Test Level E	F	G	H	Mathematics Concepts Test Level E	F	G	H
3	928*	889			877*	873		
4	916	1064*	909		801	896*	909	
5		1191	957*	945		1030	917*	876
6			985	964*			832	957*
7				982				928

*In-level sample

Table 2

Summaries of Linear Regressions of Out-of-Level

Item Difficulty Estimates on In-Level

Item Difficulty Estimates

Table 2.a - Reading Test Regression Summaries*

Grade Level Regression Order	Slope (SD)	Intercept	r
In-Level Test for Grade 3			
4 on 3	.9185 (.036)	.0001	.9644
In-Level Test for Grade 4			
3 on 4	1.0092 (.032)	.0002	.9769
5 on 4	1.1124 (.037)	.00004	.9753
In-Level Test for Grade 5			
4 on 5	.7978 (.026)	.0001	.9754
6 on 5	.9895 (.023)	-.00002	.9872
In-Level Test for Grade 6			
5 on 6	.9352 (.025)	-.00002	.9836
7 on 6	.9362 (.024)	-.00001	.9848

*(All regressions were linear with $p < .0001$.)

Table 2.b - Mathematics Concepts Test Regression Summaries*

Grade Level Regression Order	Slope (SD)	Intercept	r
In-Level Test for Grade 3			
4 on 3	.9503 (.036)	-.00004	.9666
In-Level Test for Grade 4			
3 on 4	1.0359 (.032)	-.00004	.9778
5 on 4	.9121 (.029)	.00002	.9768
In-Level Test for Grade 5			
4 on 5	.8767 (.033)	-.0001	.9671
6 on 5	.8620 (.026)	-.0007	.9791
In-Level Test for Grade 6			
5 on 6	.9874 (.030)	.00000	.9785
7 on 6	.9507 (.030)	.00004	.9763

*(All regressions were linear with $p < .0001$.)

Table 3

Summaries of Linear Regressions of Out-of-Level

on In-Level Achievement Score Estimates

Table 3.a - Reading Test Regression Summaries*

Grade Level Regression Order	In-Level Test for Grade 3 Slope (SD)	Intercept	r
4 on 3	.9929 (.001)	-.00000	.9999
Grade Level Regression Order	**In-Level Test for Grade 4 Slope (SD)**	**Intercept**	**r**
3 on 4	1.0072 (.001)	.0008	.9999
5 on 4	1.0309 (.001)	.0004	.9999
Grade Level Regression Order	**In-Level Test for Grade 5 Slope (SD)**	**Intercept**	**r**
4 on 5	.9604 (.002)	-.0023	.9999
6 on 5	1.0005 (.001)	-.0014	.9999
Grade Level Regression Order	**In-Level Test for Grade 6 Slope (SD)**	**Intercept**	**r**
5 on 6	.9897 (.001)	-.0002	.9999
7 on 6	.9897 (.001)	.0004	.9999

*(All regressions were linear with $p < .0001$.)

Table 3.b - Mathematics Concepts Test Regression Summaries*

Grade Level Regression Order	Slope (SD)	Intercept	r
In-Level Test for Grade 3			
4 on 3	.9965 (.0003)	-.0018	1.0000
In-Level Test for Grade 4			
3 on 4	1.0151 (.001)	.0010	.9999
5 on 4	.9835 (.001)	-.0006	.9999
In-Level Test for Grade 5			
4 on 5	.9760 (.001)	-.0000	.9999
6 on 5	.9706 (.001)	.0008	.9999
In-Level Test for Grade 6			
5 on 6	1.0021 (.0005)	-.0000	.9999
7 on 6	.9926 (.001)	.0002	.9999

*(All regressions were linear with $p < .0001$.)